第 43 卷 第 6A 期
2016 年 6 月
计 算 机 科 学
Computer Science
Vol. 43 No. 6A
June 2016

应用近似模糊熵的不完备信息系统属性约简

汪琼枝[1,4] 郑文曦[2] 王道然[3]
(皖西学院金融与数学学院 六安 237012)[1] (中国科学技术大学 合肥 230026)[2]
(安徽星瑞齿轮传动有限公司品控部 六安 237012)[3]
(皖西学院金融风险智能控制与预警研究中心 六安 237012)[4]

摘 要 属性约简是 Rough 集理论的重要研究内容,基于信息熵的属性约简是一种有效的属性约简方法。在实际应用中,获取的信息系统通常是不完备的。针对这种问题,在容差关系下对个体进行分类时,基于属性子集 redu 与 CAttr(属性全集)-redu 之间的内在联系,定义了一种新的知识熵,提出了一种新的应用近似模糊熵的不完备信息系统属性约简算法(newS 算法),其时间复杂度是 $O(|C|^2\sum_{i=1}^{m}(k_i^p)^2)$。最后,在 ROSE 和 UCI data 中的 6 个数据集上进行了实验仿真,结果表明 newS 算法是可行的,并且在同等约简效果下与其他算法相比具有更高的属性约简效率。

关键词 不完备信息系统,模糊熵,属性约简
中图法分类号 TP18 文献标识码 A

Attribute Reduction Algorithm for Incomplete Information Systems Based on Approximate Fuzzy Entropy

WANG Qiong-zhi[1,4] ZHENG Wen-xi[2] WANG Dao-ran[3]
(School of Finance and Mathematical, West Anhui University, Lu'an 237012, China)[1]
(University of Science and Technology of China, Hefei 230026, China)[2]
(Quaulity Control Department, Anhui Xingrui Gear-transmission Corporation, LTD, Lu'an 237012, China)[3]
(Intelligent Control and Early Warning Research Center of Financial Risk in West Anhui University, Lu'an 237012, China)[4]

Abstract Attribute reduction is important research content of rough set theory. Attribute reduction based on information entropy is an effective method of knowledge reduction. In practical application, the acquired information system is usually not complete. To solve this problem, we defined a new knowledge entropy based on the relationship between the attribute subset redu and *CAttr-redu*, and proposed a new incomplete information system attribute reduction algorithm (newS algorithm) applying approximate fuzzy entropy. Finally, simulation experiment was carried out on 6 data sets in ROSE and UCI data. The experimental results show that the newS algorithm is feasible, and has higher efficiency compared with other algorithms under the same reduction effect.

Keywords Incomplete information system, Fuzzy entropy, Attribute reduction

1 引言

Rough 集理论是一种新的处理不确定、不精确、不完备知识的软计算数学工具,在模式识别、故障诊断、机器学习等领域有广泛的应用[1-3]。信息系统中的各个属性之间相互联系、相互影响,其中也存在某些属性是冗余的。针对部分冗余属性,在保持决策分类能力不变的前提下,进行属性约简发掘并简化知识信息是 Rough 集理论的重要研究内容[4]。寻找知识信息系统最小属性约简已证明是一个 NP-hard 问题。由于设备、技术、环境等不可控因素的限制,实际问题中的知识信息数据很多都具有不确定性、模糊性、不完全性等特点,研究基于不完备信息系统的属性约简受到广泛的关注[5-7]。

近年来,国内外学者从不同的角度提出各种类型获取不完备信息系统属性约简的方法。例如,基于一般二元关系的粗糙集加权不确定性属性约简[8]、α 优势关系下不完备序信息系统的属性约简算法[9]、基于条件熵和信息熵的属性约简算法[10]、基于知识粒度的属性约简算法[11]、基于互信息大小的知识约简算法[12]等。

本文针对基于不完备信息系统容差关系下的属性约简问题,定义了一种新的知识熵,提出一种新的基于近似模糊熵的属性约简算法。在属性子集 $redu$ 下对个体进行分类时,考虑属于和不属于集合 $R_{redu}(x)$ 的个体"贡献"的不确定性的影响,将属性子集 redu 的相似类的近似度作为对属性子集整体结构的协调权重纳入模糊熵的结构中。综合考虑属性子集 $redu$ 与 $CAttr\text{-}redu$ 之间的内在的联系,定义了一种新的知识熵,提出了一种新的基于不完备信息系统下的应用近似模糊

本文受安徽省高等学校省级自然科学研究项目(KJ2013B345),安徽省高等学校省级教学研究项目(2012jyxm004),安徽省自然科学基金项目(1508085QA04)资助。
汪琼枝(1983—),女,硕士,讲师,主要研究领域为智能计算、数据挖掘,E-mail:qzwang2013@163.com;郑文曦(1982—),男,硕士,工程师,主要研究领域为智能计算、自然语言处理;王道然(1981—),男,工程师,主要研究领域为数据挖掘、故障检测。

熵的不完备信息系统属性约简算法(newS算法)。newS算法的时间复杂度是$O(|C|^2\sum_{i=1}^{m}(k_i^p)^2)$。通过在ROSE和UCI data中6个数据集上进行实验仿真,实验结果表明,本文提出的newS算法与付昂、王国胤、胡军提出的基于信息熵的不完备信息系统属性约简算法[10](IEARA算法)相比,在同等约简效果下,属性约简的效率更高。

2 相关基本概念

定义1 四元组$S=(U,C,V,f)$是一个信息系统,对于具有遗漏属性值的属性子集$P\subseteq C$,记遗漏值为"$*$",容差关系R的定义[4]为:

$$R=\{(x,y)\in U\times U\mid \forall c_i\in P\Rightarrow(c_j(x)=c_j(y)\vee c_j(x)=*\vee c_j(y)=*)\}$$

定义2 对于概率近似空间(U,R,P),系统的不确定性用系统的熵$H(R^*)$来表示[4],即

$$H(R^*)=-\sum_{i=1}^{n}P(X_i)\log P(X_i)$$

其中,$R^*=U/R=\{X_1,X_2,\cdots X_n\}$为$U$在$R$上导出的划分。

定义3 设决策表$S=(U,C\cup\{d\},V,f)$,属性集合$P\subset C$,将任意属性$c\in C/P$的属性重要度$SGF(c,P)$定义[4]为:

$$SGF(c,P)=H(P)-H(P\cup\{c\})$$

其中,$SGF(c,P)$的值越大,说明属性c对属性集合P的分类能力影响越大,即属性c对于属性集合P越重要。

定义4 设决策信息系统$S=(U,C\cup\{d\},V,f)$,条件属性集C的熵为$H(C)$。称属性集$P\subseteq C$是决策信息系统S的一个熵约简[4],当且仅当$H(P)\leqslant H(C)$,且对于任意的属性子集$P'\subset P$,都有$H(P')>H(C)$成立。

类似完备信息系统,不完备信息系统的约简集一定存在,但未必是唯一的。

定义5 在给定的二元关系R下,对象$x\in U$关于属性子集P的相似类记为$R_P(x)=\{y\in U\mid R(x,y)\}$,定义$r(R_p(x))=\frac{|R_p(x)|}{|U|}$是对象$x\in U$关于属性集$P$的相似类的近似度。

3 应用近似模糊熵的不完备信息系统属性约简算法

相对于属性全集$CAttr$,属性子集$redu$与$CAttr-redu$是一个整体的两个部分,两部分对个体的分类影响在一定条件下相互影响、相互制约,基于属性子集$redu$与$CAttr-redu$之间的内在的联系,定义了一种新的知识熵。

3.1 应用近似度的模糊熵

定义6 决策信息系统$S=(U,C\cup\{d\},V,f)$,其中U为一个非空的有限对象集,C是条件属性集,$\{d\}$是决策属性集,对每个属性$c\in C\cup\{d\}$,R为信息函数,属性集合$P\subseteq C$。定义P的应用近似度的模糊熵为:

$$H(P)=\sum_{x\in U}r(R_p(x))d_z(F_{R_p(x)})(1-d_z(F_{R_p(x)}))$$

其中$d_z(F_X)=-\frac{1}{|U|}\sum_{j=1}^{m}p_j\ln p_j+(1-p_j)\ln(1-p_j)$,$d_z(F_X)$是$X\subseteq U$基于信息观下粗糙集的模糊度度量,$p_j=k_j/|X|$,$k_j$为集合$X$中决策属性值为$v_j^d$的实例个数,$|X|$为集合$X$的基数。

$H(P)\geqslant 0$,$H(P)$反映了属性子集P在论域U上对信息系统S的分类混乱程度。若$H(P)$越小,则说明分类能力越强;若$H(P)$越大,则说明分类能力越弱;当$H(P)=0$时说明分类清晰。

$H(P)$的定义由3部分构成:

1)$r(R_p(x))$是关于属性子集P的相似类的近似度;

2)$redu$与$CAttr-redu$是属性全集的两部分,彼此相互影响、相互制约;

3)采用$d_z(F_{R_p(x)})$和$(1-d_z(F_{R_p(x)}))$均衡度量,共同刻画粗糙集的不确定性,模糊性更合理。

3.2 应用近似模糊熵的不完备信息系统属性约简算法

在应用近似度的信息熵的定义下,以空集为不完备信息系统属性约简的起点,按照单个属性的属性重要性大小,由大到小逐个加入约简集合,直到约简集合的模糊熵小于等于原属性集的模糊熵,再按照$redu$中各属性的加入顺序,逐个检查属性c是否满足$SGF(c,redu/\{c\})\leqslant 0$,删去满足条件的属性,得到容差关系下的不完备信息系统的约简。

在相容关系下,应用近似模糊熵的不完备信息系统属性约简算法(newS算法)如下。

输入:一个决策信息系统为S=(U,C∪{d},V,f)。

输出:该决策信息系统的一个约简redu。

Step1:初始化:redu=∅,CAttr=C,d,e=H(C)。

Step2:计算H(redu),若H(redu)>e,则

Step 2.1:计算研究对象$x\in U$关于属性集redu的相似类的近似度

$$r(R_{redu}(x))=\frac{|R_{redu}(x)|}{|U|}$$

Step 2.2:计算集合$R_{redu}(x)\subseteq U$基于信息观下粗糙集的模糊度度量

$$d_z(F_{R_{redu}(x)})=-\frac{1}{|U|}\sum_{j=1}^{m}p_j\ln p_j+(1-p_j)\ln(1-p_j)$$

Step 2.3:计算属性集redu的应用近似度的模糊熵

$$H(redu)=\sum_{x\in redu}r(R_{redu}(x))d_z(F_{R_{redu}(x)})(1-d_z(F_{R_{redu}(x)}))$$

Step 2.4:针对每个条件属性$c_i\in CAttr$,计算$H(redu\cup\{c_i\})$;

Step 2.5:根据定义2,计算每个条件属性$c_i\in CAttr$的属性重要度:

$$SGF(c_i,redu)=H(redu)-H(redu\cup\{c_i\})$$

Step 2.6:在集合CAttr-redu中找到属性c_j,使得

$$c_j=\underset{CAttr}{ArgMax}\{SGF(c_i,redu)\}$$

Step 2.7:执行$redu=redu\cup\{c_j\}$,$CAttr=CAttr/\{c_j\}$,并记录redu中各属性加入的顺序;

Step3:若$H(redu)\leqslant e$,则按加入到redu的属性的顺序,从后至前逐个检查每个属性:

$SGF(c,redu/\{c\})\leqslant 0,(c\in redu)$

若成立,则$redu=redu/\{c\}$,否则$redu=redu$

Step4:算法结束。

注:在Step2.7中,若有多个属性c_j符合条件,则各属性c_j分别执行$redu=redu\cup\{c_j\}$,$CAttr=CAttr/\{c_j\}$操作,记录顺序时,按照c_j下标顺序排列次序。

在实际决策表中,条件属性C的不同属性值个数m通常情况下有$m>1$,因此有

$$O(\sum_{i=1}^{m}|C|(k_i^p)^2)<O(|C||U|^2)$$

所以,newS算法的时间复杂度为:$O(|C|^2\sum_{i=1}^{m}(k_i^p)^2)$。

4 仿真实验

用ROSE和UCI data中6个数据集(见表1)在容差关系

(下转第102页)

根据上述预测步骤，并取 $m=1000$，预测 1984 年的 4 个季度的 GDP，预测结果与相对误差如表 2 所列。

表 2　1984 年 4 个季度的 GDP 预测结果的相对误差

时间	算法 3 误差（出口额）	算法 3 误差（人均收入）	算法 4 误差	算法 5 误差
1984-I	0.0092	0.0038	0.0052	0.0010
1984-II	0.0052	0.0060	0.0079	0.0035
1984-III	0.0014	0.0103	0.0120	0.0079
1984-IV	0.0017	0.0032	0.0055	0.0029
2 范数	0.0108	0.0130	0.0162	0.0092

由表 2 可见，本文提出的基于相依条件云的二维云推理方法（算法 5）的预测效果优于传统的双条件单规则（算法 4）、单条件单规则（算法 3）发生器算法。这也表明本文提出算法处理具有相依关系二维云推理是可行、有效的，这也弥补了传统双条件单规则推理的不足。

参 考 文 献

[1] 邸凯昌，李德毅，李德仁. 云理论及其在空间数据挖掘和知识发现中的应用[J]. 中国图像图形学报，1999，4(11)：930-935

[2] 杨朝晖，李德毅. 二维云模型及其在预测中的应用[J]. 计算机学报，1998，21(11)：961-969

[3] 李德毅，杜鹢. 不确定性人工智能[M]. 北京：国防工业出版社，2005：335-361

[4] 陈昊，李兵. 云推理方法及其在预测中的应用[J]. 计算机科学，2011，38(7)：209-224

[5] 蒋嵘，李德毅，陈晖. 基于云模型的时间序列预测[J]. 解放军理工大学学报，2000，1(5)：13-18

[6] 张仕斌，许春香. 基于云模型的信任评估方法研究[J]. 计算机学报，2013，36(2)：422-431

[7] 李德毅，史雪梅，孟海军. 隶属云和隶属云发生器[J]. 计算机研究和发展，1995，32(6)：15-20

[8] 李德毅，刘常昱. 论正态云模型的普适性[J]. 中国工程科学，2004，6(8)：28-33

[9] Chen Hao, Li Bing. Approch to uncertain reasoning based on cloud model[J]. Journal of Chinese Computer Systems, 2011, 32(12): 2449-2455

（上接第 82 页）

下对文献[10]中的 IEARA 算法和本文中的 newS 算法进行了分析比较（见表 2）。选用 CPU：AMD Athlon64(2800+)，1.0GB 内存配置的计算机进行仿真实验。

表 1　测试数据集信息

序号	数据集名称	数据集来源	样本容量	C	完备
1	Breast cancer wisconsin	UCI	699	9	否
2	Primary-tumor	UCI	339	17	否
3	Audiology-standardized	UCI	200	69	否
4	Crx-local	ROSE	690	15	否
5	House-votes-84data	ROSE	435	16	否
6	Lsd265	ROSE	265	35	是

表 2　基于容差关系的属性约简算法比较(ms)

序号	IEARA 算法约简结果	IEARA 算法约简时间	newS 约简结果	newS 约简时间	IEARA-newS
1	5	25781	5	22937	2844
2	16	44110	16	38000	6110
3	16	72359	16	62110	10249
4	13	68453	13	55484	12969
5	15	46578	15	38204	8374
6	9	30078	9	25312	4766

将两种算法约简时间进行绘图比较，可以更加直观地比较两个算法的效率关系，如图 1 所示。

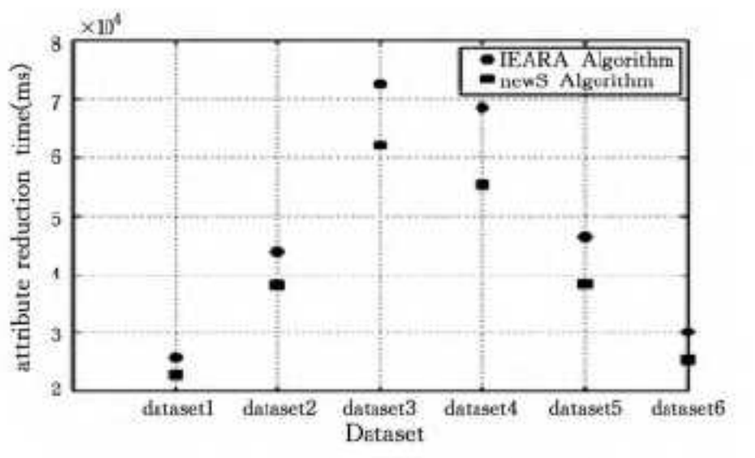

图 1　IEARA 算法与 newS 算法比较

从图中可以看出，newS 算法的约简时间总是低于 IEARA 算法的约简时间。

结束语　本文研究基于容差关系下的不完备信息系统属性约简算法。基于属性约简子集 $redu$ 与属性全集 $CAttr$ 之间的关系，将属性子集 $redu$ 的相似类的近似度作为对属性子集整体结构的协调权重纳入模糊熵的结构中，综合考虑属性子集 $redu$ 与 $CAttr$-$redu$ 之间必然具有的内在的联系，定义了一种新的知识熵，提出了一种新的应用近似模糊熵的不完备信息系统属性约简算法。在 ROSE 和 UCI data 数据集上进行仿真验，实验结果表明本文提出的 newS 算法不仅是可行的，而且具有较高的属性约简时间效率。

参 考 文 献

[1] 戴逸翔，王雪，李宣平，等. 面向生物信息感知网络稀疏脑电测量的模糊粗糙情绪识别[J]. 仪器仪表学报，2014，35(8)：1693-1698

[2] 韩利强，陈泽华，曹长青，等. TEP 故障诊断方法研究[J]. 计算机应用与软件，2014，31(7)：82-85

[3] 马文萍，黄媛媛，李豪，等. 基于粗糙集与差分免疫模糊聚类算法的图像分割[J]. 软件学报，2014，25(11)：2675-2689

[4] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2008

[5] 曾晓辉，文展. 不完备信息系统的属性约简算法[J]. 计算机工程，2009，35(24)：185-187

[6] 滕书华，周石琳，孙即祥，等. 基于条件熵的不完备信息系统属性约简算法[J]. 国防科技大学学报，2010，32(1)：90-94

[7] 周志平，刘付显. 一种改进的模糊粗糙集知识约简方法[J]. 计算机工程与应用，2012，48(18)：132-135

[8] 滕书华，鲁敏，杨阿锋，等. 基于一般二元关系的粗糙集加权不确定性度量[J]. 计算机学报，2014，37(3)：649-665

[9] 韦碧鹏，吕跃进，李金海. α 优势关系下粗糙集模型的属性约简[J]. 智能系统学报，2014，9(2)：251-258

[10] 付昂，王国胤，胡军. 基于信息熵的不完备信息系统属性约简算法[J]. 重庆邮电大学学报，2008，20(5)：586-592

[11] 李秀红，史开泉. 一种基于知识粒度的不完备信息系统的属性约简算法[J]. 计算机科学，2006，33(11)：169-170，199

[12] 王青海. 互信息的序决策信息系统属性约简研究[J]. 计算机工程与设计，2012，37(7)：2822-2826